Running head: Quantification of inter-test variability in ecotoxicity data

Corresponding author:

Dr. Peter S. Craig

Department of Mathematical Sciences
Durham University
Science Laboratories
South Road
Durham
County Durham
DH1 3LE
United Kingdom

Tel: +44 (0)191 334 3050
Fax: +44 (0)191 334 3051

Total number of words: 7000 (approx.)

On the quantification of inter-test variability in ecotoxicity data with application to species sensitivity distributions

Graeme L. Hickey[†], Peter S. Craig[‡], Robert Luttik[#] and Dick De Zwart[#]

[†]*Northwest Institute of BioHealth Informatics, Manchester University, Manchester, UK*

[‡]*Department of Mathematical Sciences, Durham University, Durham, UK*

[#]*RIVM, Bilthoven, The Netherlands*

29 *To whom correspondence may be addressed: P.S.Craig@durham.ac.uk.

Abstract

Ecotoxicological hazard assessment relies on species effect data to estimate quantities such as the predicted no-effect concentration. Whilst there is a concerted effort to quantify uncertainty in risk assessments, the uncertainty due to inter-test variability in species effect measurements is an overlooked component. The EU REACH guidance document suggests that multiple toxicity records for a given chemical-species combination should be aggregated by the geometric mean. Ignoring this issue or applying unjustified so-called harmonisation methods weakens the defensibility of uncertainty quantification and interpretation about properties of ecological models, for example the predicted no-effect concentration.

In the present study we propose a simple and broadly theoretically justifiable model to quantify inter-test variability and analyse it using Bayesian methods. The value of data in ecotoxicity databases is maximised by utilising (interval-)censored data. An exploratory analysis is provided to support the model. We conclude, based on a large ecotoxicity database of acute-effects to aquatic species, that the standard deviation of inter-test variability is about a factor (or fold-difference) of 3. The consequences for decision makers of (not) adjusting for inter-test variability are demonstrated.

Keywords: Inter-test variability, REACH, Species sensitivity distribution, Toxicity data, Bayesian statistics

1 Introduction

A fundamental component of ecotoxicological risk assessment within the European Union (EU) regulation concerning the 'Registration, Evaluation, Authorisation & restriction of CHemicals', better known as REACH', is the *predicted no-effect concentration* (PNEC) [1]. This is divided by an estimate of the *predicted environmental concentration* (PEC) to yield the *risk characterisation ratio* (RCR). Risk assessment frameworks of chemical products and water quality criteria which do not fall under the remit of REACH (e.g. pesticides [2] and metals [3, 4]) or those outside of the EU (e.g. United States [5, 6, 7], Canada [8] and Australia and New Zealand [9]) also generally rely on quantities equivalent to the PNEC, although terminology and the mechanics do differ slightly.

All of the aforementioned frameworks rely on ecotoxicity data at some tier, whether through the application of assessment factors [10] or probabilistic modelling (e.g. species sensitivity distributions [11]). Standard types of ecotoxicity data are (i) concentrations which affect x% of members of a species with respect to a given toxicological endpoint (ECx; denoted LCx when the endpoint is lethality) and (ii) no-observed effect concentrations (NOECs). As an example, the current REACH guidance document (GD) [12] defines $\text{PNEC}_{\text{aquatic}}$ for freshwater compartments according to one of two methods. The first is the minimum observed toxicity value divided by an assessment factor between 1000 and 10 which is determined according to the type, quantity and taxonomic diversity of the measured toxicity data. The second is as the estimated 5-th percentile of a species sensitivity distribution (SSD) which is fitted to a minimum of 10 NOEC values (spanning a minimum level of taxonomic diversity), called the hazardous concentration to 5% of species (HC5), and subsequently divided by an SSD-specific assessment factor between 5 and 1. In principle, an SSD can be fitted to acute toxicity data and extrapolated *a posteriori* using an acute-to-chronic assessment factor; this approach is not currently endorsed by REACH.

The general SSD model only describes the interspecies variability. Some practitioners [13, 14, 15] have also incorporated sampling variation and assemblage parameter uncertainty into estimation methods. The actual uncertainty about the toxicity values used to fit the SSD and derive a PNEC (or similar quantity) is generally overlooked [16]. In this regard, we shall use the term *inter-test variability* to refer to variability, potential or actual, between test results for the same chemical on the same species. Inter-test variability is implicitly acknowledged in the REACH GD [12] under the description of the sources of uncertainty intended to be accounted for by assessment factors. Inter-test variability is present even when only a single ECx or NOEC is a available for a particular chemical-species combination but we observe it empirically when, for a given chemical-species combination, there exist *multiple* records (i.e. toxicity values) that are considered to be broadly comparable (e.g. all acute median effect concentrations), a situation often found in the analysis

of large databases of existing data.

Inter-test variability has several sources, including: (1) inter- and intra-laboratory variation; (2) intraspecies variation (biological variance); (3) variation in experimental conditions (e.g. pH, salinity, water hardness, formulation); and (4) differences in dose-response modelling and statistical analysis. There is considerable overlap between (1) and the others. The European Food Safety Authority (EFSA) has referred to *measurement uncertainty* as the first two items [17].

We make no blanket definition of inter-test variability as it would require a judgement as to which particular potential sources of such variability would be considered relevant or acceptable in any particular context. Nonetheless, in choosing to pool certain data in a statistical analysis or to average certain records for risk assessment, a judgement is being made. For the example we give later, the scope of inter-test variability is defined clearly by our choice of rules for selecting records from a database.

A further potential source of inter-test variability is variation in the effect endpoint measured. For example, if three EC50 values are available for Chlorpyrifos tested on *Daphnia magna* with effects on mortality, growth and reproduction, aggregating them into a single EC50 measurement to be applied in a risk assessment incorporates this addidional source into the inter-test variability. Combining concentrations for different endpoints may be controversial but is part of some current practice; a research database developed by the United States Environmental Protection Agency (US EPA) to build interspecies correlation estimation models [18] aggregates acute lethal (i.e. LC50) and *sub*-lethal effects (EC50 for immobilisation). If more than one life-stage is of interest for a species, this too may become a component of inter-test variability. When considering a large database, there may be a temporal element to laboratory variation since analytical techniques have improved over the years.

There is relatively little available information quantifying overall inter-test variability; this is in contrast to (i) the wealth of data quantifying one component, namely statistical uncertainty for the specific dose-response model used to analyse the data, which is published alongside most effect concentrations; and (ii) chemical-specific studies of inter- and intra-laboratory variation. Baird et al. [19] note that standardised laboratory toxicity tests performed with *D. magna* and the same chemical may vary by a factor of 2–3 within and between laboratories. Raimondo et al. [20] notes that in the development of the aforementioned US EPA research database, the percentage of records for a chemical-species combination differing by less than a factor of 2, 5, and 10 was 56%, 86%, 94% respectively. Fairbrother [21] reports that differences between records for a chemical-(aquatic-)species combination can be as great as a factor of 5; a similar difference was recorded for a US EPA wildlife toxicity research database [22].

Existing probabilistic risk assessment frameworks do not address inter-test variability quantitatively since they rely on the species effect data to be precisely known. In the REACH GD [12, pp. 7–8, 21–22] it is

required that records for a chemical-species combination are aggregated; this is so-called 'harmonisation' . The procedure can be considered as a type of meta-analysis. For each chemical-species combination: (1) filter the data measurements according to a systematic review of the reported experimental conditions; (2) if multiple values remain, test if the maximum value exceeds the minimum value by more than a single order of magnitude; (3) take the geometric mean and apply as a substitute value. If the outcome of step (2) is greater than a single order of magnitude, further review is required. The International Council on Mining and Metals (ICMM) [4] provide similar guidance with respect to metal toxicity, stating additionally that normalisation may be appropriate if differences in values are an apparent result of differences in bioavailability.

Current SSD practice does not take the presence of inter-test variability into account when fitting SSDs and estimating hazardous concentrations. The ICMM states that the focus of risk assessment "should be on interspecies variability and not on intraspecies variability" [4]. This requirement in conjunction with the REACH requirement of transparent uncertainty analysis [23] motivates the present study. Simply fitting an SSD to individual or aggregated estimated effect concentrations includes inter-test variability along with interspecies variation in the SSD; consequently this undermines the interpretation of the estimated hazardous concentration as a summary of interspecies variation. In the present study, we first model and quantify inter-test variability using a large database of ecotoxicity data for aquatic species and then consider the effect of taking the magnitude of inter-test variability into account when estimating the hazardous concentration as a summary of interspecies variation. Although the consequences of inter-test variability under the scope of REACH will stem from chronic data, we use acute data because there exist much larger databases of acute data to analyse. Quality-controlled metadata (e.g. experimental conditions, life-stage) is unavailable for many records; therefore, we do not explore individual components of inter-test variability.

2 Data and methods

2.1 Data

A large aquatic ecotoxicity research database[24] was used which is comprised of 30,369 acute (EC50 and LC50) and chronic (NOEC) records spanning 3442 distinct chemicals and 1549 species. Approximately 22,000 records were extracted from the U.S. EPA ECOTOX database; the remainder were extracted from multiple other U.S. EPA and RIVM programme databases; all references are available in De Zwart [24], Section 8.2.1. Key fields of the database include: species, chemical, endpoint, effect, duration of experiment, whether the endpoint was acute or chronic (denoted A/C), concentration (μg/L) and whether the measurement was censored or pointwise. Incomplete experimental data were available for some records; see De Zwart [24] for

further details. The database is freely available as Supporting Information.

In addition to scientific review of original data sources, De Zwart [24], pp. 136–138 describes an *ad hoc* collection of data filtering queries applied to a larger database which yielded the research database described here. In particular, censored data points were removed unless they were either the smallest, greatest, or only reported concentration for the corresponding chemical-species-A/C combination. Therefore, it is likely that some records, whether 'outliers' or censored values, have been removed which would have been informative for making inferences about inter-test variability.

We predominantly focus on a subset of the database selected according to the following rules: (i) all records are either LC50 or sub-lethal EC50 (effect defined as immobility) values; (ii) each record is identifiable at the species level; (iii) species belonging to the taxonomic order *Insecta* or *Crustacea* must have a minimum 48h exposure; (iv) species not belonging to the taxonomic order *Insecta* or *Crustacea* must have a minimum 96h exposure; (v) no qualitative 'approximate' values were admissible; and (vi) each chemical must have at least 5 distinct species pointwise measurements. Acute and chronic data are not amalgamated for purposes of estimating inter-test variability since it would be a source of systematic error. Item (v) enhances comparability of results reported here of statistical analyses which have varying data requirements; this is described in Section 2.4. In this data subset, there are 6576 records: 6279 classed as pointwise; 112 as interval censored; 173 as right censored and 12 as left censored. Of these 6576 records there were 4854 unique chemical-species combinations spanning 339 chemicals and 610 species.

We will use the following notation. Let y_{ijk} be the k-th log (base 10) transformed toxicity value for species j tested on chemical i. Also, let K_{ij} be the number of toxicity records for species j tested on chemical i. The three possible cases are: (1) $K_{ij} = 0$ which means zero records are available for chemical-species combination (i, j); (2) $K_{ij} = 1$ which means precisely one record is available; and (3) $K_{ij} > 1$ which means multiple records are available. It is case (3) which allows for inferences to be made about inter-test variability; when it is part of a larger statistical model, case (2) data can also influence model parameter estimates.

2.2 Exploratory analysis

The most straightforward exploratory analysis is to calculate the sample standard deviation of log toxicity values for each chemical-species combination (i, j) for cases when $K_{ij} > 1$, namely

$$s_{ij} = \sqrt{\frac{1}{K_{ij} - 1} \sum_{k=1}^{K_{ij}} (y_{ijk} - \bar{y}_{ij})^2},$$

where $\bar{y}_{ij}$ is the sample mean for chemical-species combination (i, j). Whilst this statistic is non-parametric, it is only calculable with pointwise toxicity data. Therefore, we only analyse the 6279 pointwise records extracted previously.

2.3 Classical modelling

Inferences about a simple exploratory analysis can be made if a model is proposed. Pragmatic modelling is advocated here in light of the limited number of records for chemical-species combinations generally available in ecotoxicity databases and the lack of quality controlled metadata. We therefore propose a simple model, namely

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk},$$

where μ_{ij} is the 'true' log-transformed toxicity value for chemical-species combination (i, j) and ε_{ijk} is the inter-test variability. In addition, we make two initial modelling assumptions: (i) inter-test variability in the database subset is random and not systematic, thus not requiring bias correction; and (ii) $\varepsilon_{ijk} \,|\, \sigma \sim N(0, \sigma^2)$, where the tilde is read as 'distributed'. These two assumptions combine to state that each residual about the 'true' log-transformed toxicity value for chemical-species combination (i, j) is a random sample from a normal distribution centered about zero with homogeneous variance σ^2 that is independent of experimental conditions, chemical and species.

Since there is no *a priori* reason why the sum of variation, which is extraneous to the interspecies variation, should be a unique property of the specific risk assessment (i.e. chemical or species tested) rather than one which is globally defined, this modelling assumption appears reasonable. This model is, however, incompatible with the joint modelling of multiple arbitrary endpoints, e.g. acute and chronic together. Although the normal distribution assumption allows for confidence statements to be made, testing the assumption is difficult since for each chemical-species combination (i, j), K_{ij} — the number of available L(E)C50 measurements for the chemical-species combination, is generally too small to (confidently) make inferences from goodness-of-tests. This exact problem is faced by risk assessors trying to fit SSDs. Despite there being strong criticism of distributional assumptions [25], the REACH GD [12] states that normality is:

> 'the pragmatic choice... because of the available description of its mathematical properties (methods exist that allow for most in depth analyses of various uncertainties)'.

It is also a standard distributional choice in statistical modelling of errors [26], therefore we adopt it here. If inter-test variability can be envisaged as the sum of many smaller error components, then appealing to the central limit theorem [26] would offer some further justification for the model proposed here.

Based on these assumptions, a statistically unbiased estimate of σ^2 is the pooled variance, namely

$$s^2_{\text{pooled}} = \frac{\sum_{(i,j)}(K_{ij}-1)s^2_{ij}}{\sum_{(i,j)}(K_{ij}-1)}, \tag{1}$$

where the sum is over all chemical-species combinations (i, j) in the database. Pooling variances across chemicals is not new [27, 28]; however, here we are pooling inter-test variances rather than interspecies variances. For each K_{ij} bin we can determine whether the variation in observed s_{ij} values is consistent with distributional assumptions, or whether there is over- or under-dispersion. An approximate 95% interval for each bin $K_{ij} = K^*$ is determined as $\left(\chi_{K^*,0.025}\hat{\sigma}/\sqrt{K^*-1}, \chi_{K^*,0.975}\hat{\sigma}/\sqrt{K^*-1}\right)$ with $\hat{\sigma} = s_{\text{pooled}}$, where $\chi_{(K^*-1),\alpha}$ is the 100α-th percentile of the Chi distribution with K^*-1 degrees of freedom.

2.4 Bayesian modelling

The application of Bayesian modelling in ecotoxicological risk assessment has recently gained attention [13, 10, 29, 14, 36, 30, 31, 32, 33, 15]. The basic idea of the Bayesian paradigm is that prior knowledge (or some suitable objective proxy) can be updated with the data likelihood function to yield a posterior distribution of the unknown model parameters, from which probabilistic statements can be made. For an introduction to Bayesian methods in ecological risk assessment, consult Warren-Hicks and Hart [34] and references therein.

We earlier proposed the data model as $y_{ijk} \,|\, \mu_{ij}, \sigma \sim N(\mu_{ij}, \sigma^2)$ for $k = 1, \ldots, K_{ij}$ and species j tested on with chemical i; this defines the *likelihood function* for a given ecotoxicity database. In order to analyse the model from a Bayesian perspective we need to specify [prior] distributions for (i) σ^2 and (ii) μ_{ij}.

For (i), a default prior distribution, which is referred to as *non-informative*, is the Jeffreys prior: $\pi(\sigma^2) \propto \sigma^{-2}$ for $\sigma^2 > 0$. This is equivalent to the assumption that all values of $\log(\sigma)$ are equally likely *a priori*. Although not universally accepted, it has been applied to the interspecies standard deviation parameter in many Bayesian SSD analyses [13, 35, 10, 36, 15]. See Gelman [37], pp. 62–66 and references therein for general consideration of Jeffreys' and other priors.

For (ii), the standard ecotoxicological model is the normal species sensitivity distribution, which can be described in terms of a probability distribution, namely

$$\mu_{ij} \,|\, \alpha_i, \psi_i \sim N(\alpha_i, \psi_i^2), \tag{2}$$

where α_i and ψ_i are the per-chemical SSD mean and standard deviation parameters on log concentration for chemical i. Note that, in a situation where ecotoxicity data are measured for a single chemical only

without inter-test variability, i.e. $\sigma = 0$, the model reduces to the one described in Aldenberg and Jaworska [13] subject to notational differences; their μ and σ are then our α_i and ψ_i respectively and our μ_{ij} then correspond to their data. The inclusion of a common inter-test variability parameter means that the SSDs are only conditionally independent (probabilistically) between chemicals. This hierarchical structure requires us to either estimate each pair of hyper-parameters (α_i, ψ_i) or model them; we do the latter using standard independent non-informative priors: $\pi(\alpha_i, \psi_i) \propto 1$. This is equivalent to the assumption that all values of $\psi_i > 0$ are equally likely *a priori*. The Jeffreys prior used for the inter-test variance parameter is not applicable here since it would lead to an unbounded integrated probability density; see [37], pp. 521–522. The approach taken here was made to meet the necessary technical requirements and is deemed reasonable by Gelman [37]. Alternatively, one may consider the per-SSD parameters as *exchangeable* [10, 15] and model them as coming from a larger hyper-population, or use expert elicitation to specify prior beliefs. This, however, is beyond the scope of the present study which is primarily interested in estimating σ and in the consequences within the remit of the current simple SSD modelling framework.

The requirement in the database subset extraction routine that each chemical must have a minimum of 5 distinct species with pointwise measurements was made to overcome a technical issue regarding the analytical structure of the posterior distribution and to reduce sensitivity to the choice of prior distribution for the ψ_i parameters. Consult the Supporting Information for further details.

The posterior distribution of the parameters of interest is calculated using Bayes' rule and is proportional to

$$\prod_{i=1}^{N} \prod_{j \in J_i} \prod_{k=1}^{K_{ij}} \ell(y_{ijk} \mid \mu_{ij}, \sigma^2) \pi(\mu_{ij} \mid \alpha_i, \psi_i) \pi(\alpha_i) \pi(\psi_i) \pi(\sigma^2), \tag{3}$$

where J_i is the set of all species tested with chemical i. A mathematical derivation of the posterior distribution, and a description of the sampling methodology used to analyse it and a computer code script for running it are all given in the Supporting Information.

The normal distribution assumption is not a prerequisite of this analysis; alternative distribution proposals for SSDs include the logistic [38]; Burr Type III [39]; triangular [5]; and uniform, exponential and Weibull [40]. This is an ongoing and widely debated issue in the ecotoxicological risk assessment arena. We adopt the normal distribution based on its prevalence in the ecotoxicological SSD-based risk assessment arena and its convenient properties for mathematical analysis of the posterior distribution.

2.5 Consequences in setting environmental standards

Inter-test variability is not taken into account in standard procedures for determining hazardous concentrations, although it should be due to the simple fact that it is an additional component of uncertainty.

We briefly examine the consequences of adjusting for its presence. The usual method for fitting an SSD assumes the 'true' log-transformed toxicity value μ_{ij} for chemical-species combination (i, j) to be equal to the aggregated measurements $\bar{y}_{ij}$; no account of uncertainty is made. Therefore, we define the 'inter-test variability adjusted' (ITVA) SSD for chemical i to be the distribution of the μ_{ij} values, the true interspecies variation. From this ITVA SSD, we can in turn calculate a ITVA estimate of the HC5 which can be used to set environmental standards. A numerically simple method is to observe that based on the model described by Equation. 2, the ITVA $\log_{10}$(HC5) for chemical i, which extrapolates for interspecies variation only, is equal to $\alpha_i - K_5\psi_i$ where K_5 is the 95-th percentile of the standard normal distribution [13]. The posterior distribution of the ITVA $\log_{10}$(HC5) can be calculated from the full posterior distribution (Eqn. 3).

A comparison between the ITVA HC5 estimator and the usual HC5 estimator (i.e. ignoring the issue of inter-test variability by aggregating multiple chemical-species combination) allows us to infer the consequences of accounting for inter-test variability in ecotoxicological risk assessment. For the Bayesian analysis, the median of the ITVA $\log_{10}$(HC5) distribution is calculated for each chemical in the robust database subset for (a) all the data, and (b) the pointwise-only data. The latter allows for direct comparability with the usual estimator methodology which assumes pointwise data. Details of how the Bayesian hierarchical model can be fit with censored data is described in the Supporting Information. For the usual method, we calculate the median $\log_{10}$(HC5) for each chemical in the robust database subset, using the frequentist method described in Aldenberg and Jaworska [13] who also showed that the estimator corresponded to the Bayesian posterior median under Jeffreys prior [36]; we used pointwise-only toxicity data, harmonising multiple measurements for the same chemical and species using the geometric mean, as per the guidance in the REACH GD [12, pp. 7–8]. Qualitative aspects of the meta-analysis aggregation method were not undertaken as the database was developed to meet strict quality control standards [24] *a priori*, thus residual variation is scientifically attributable to inter-test variability. Moreover, we consider the REACH GD threshold of one order of magnitude to be entirely arbitrary.

3 Results

3.1 Empirical & classical analysis

In **Figure 1** we display boxplots, for the pointwise-only subset of the acute data, of s_{ij} for every chemical-species combination (i, j) for which $K_{ij} > 1$ (s_{ij} is not defined otherwise). Note that s_{ij} is reported in units of $\log_{10}$ μg/L. Since sampling uncertainty will undoubtedly be greater for small K_{ij}, we stratify the boxplots according to K_{ij} bins. The red dashed line shows the estimated pooled standard deviation, $s_{\text{pooled}} = 0.507$.

Since s_{ij} is measured on the log (base 10) scale, s_{pooled} corresponds to a factor (or fold-difference) of about 3.2. There is no evidence of s_{ij} being explained by major taxonomic grouping, however this conclusion is not surprising since 79% of the 4615 distinct (i, j) combinations are for *Crustacea* and *Osteichthyes* only — a reflection of the imbalance between ecotoxicity databases and ecological representativeness.

Conditional on the model assumption that the log-toxicity values are realisations from a normal distribution with homogeneous variance (with the mean equal to the 'true' log-toxicity value for that chemical-species combination), an approximate 'region of high probability' for future s_{ij} is highlighted blue in Figure 1. For $2 \leq K_{ij} \leq 6$ the median of the boxplots tend towards s_{pooled} from below which, assuming the population inter-test variance is homogenous, is expected since the sampling distribution of s_{ij} is skewed to the right. As shown by the grey line graph overlay, the number of chemical-species combinations for which K_{ij} is large ($K_{ij} \geq 8$) is generally less than 5. Although as K_{ij} increases the standard error about s_{ij} reduces, with only a handful of (i, j) combinations there will be less information to gauge whether homogeneity is a reasonable assumption. Qualitatively we conclude that homogeneity is a reasonable hypothesis. The null hypothesis of homogeneity was examined using Levene's test based on the sum of *median* squares. A P-value of 0.656 ($F = 0.8125$ on 14 and 912 degrees of freedom) does not provide evidence to reject the hypothesis of homogeneity. This test, however, is limited in value since it is not generally appropriate to consider meaningful populations (in the statistical sense) defined by the K_{ij} bins.

As a side analysis to the focal *acute* dataset in the present study, an estimate of the NOEC-based inter-test pooled standard deviation is calculated as 0.580. There are much fewer data available (174 distinct (i, j) combinations with repeated measurements such that 145 combinations fall into the $K_{ij} = 2$ bin; 17 in the $K_{ij} = 3$ bin; 8 into the $K_{ij} = 4$ bin; 3 into the $K_{ij} = 5$ bin and 1 into the $K_{ij} = 6$ bin) to test the homogeneity hypothesis, although qualitative analysis suggested the hypothesis was reasonable. Although test methods for assessing chronic (NOEC) toxicity are inherently more complex than those for assessing acute (EC50 / LC50) toxicity, conditional on the homogeneity model, the average increase in inter-test variability is only 18%. However, the uncertainty about this estimate is also larger.

3.2 Bayesian analysis & consequences

Using Markov chain Monte Carlo sampling methods, 10000 samples were drawn from the posterior distribution (Eqn. 3); technical details of this are provided in the Supporting Information. A kernel density plot of σ (derived by applying the square-root function to all samples of σ^2) is shown in **Figure 2**. The posterior median of σ is 0.466 $\log_{10}$ μg/L with 95% credible interval (the Bayesian analogue of a confidence interval) $(0.454, 0.480)$. The frequentist estimate, s_{pooled}, falls outside the Bayesian credible interval; by

fitting some additional models to the data, the slight difference between s_{pooled} and the posterior median estimate, which is of negligible practical significance, was found to be due largely to the hierarchical (SSD) modelling of the chemical-species mean toxicities μ_{ij} rather than to any difference between Bayesian and frequentist procedures or to the incorporation of censored data in the Bayesian analysis. The posterior distribution of σ was also found to be insensitive to the choice of prior distribution for σ.

In **Figure 3**, for 339 chemicals we plot the posterior ITVA median $\log_{10}$(HC5) estimate (based on all data [left panel] and pointwise-only data [right panel]) against the usual $\log_{10}$(HC5) estimate which aggregates measurements by the geometric mean. There is a strong linear correlation between the two estimates. The median difference between the estimates was 0.152 and 0.157 based on the inclusion and omission of censored data respectively. This corresponds to 83% of posterior median estimates being *larger* than the usual estimates, i.e. accounting for ITV leads to a more conservative estimator on average. The standard deviation and 95% quantile interval of the difference between ITVA and usual $\log_{10}$(HC5) estimates were respectively 0.192 and $(-0.264, 0.469)$ when censored data was included; and 0.160 and $(-0.166, 0.443)$ when censored data was omitted.

4 Discussion

Inter-test variability is a source of uncertainty and therefore *should* be considered by risk assessors and risk managers. Since uncertainty analysis is a necessary requirement under the REACH GD at the intermediate and higher tiers of risk assessment [23], it seems contradictory that the same GD authorises averaging out the effects of inter-test variability [12].

According to [11], the HCx is the concentration hazardous to x% of species; equivalently, the probability that a randomly selected species from the assemblage has its endpoint exceed is x%. It can be inferred from common practice and risk assessment guidance that the SSD, of which the HCx is a summary statistic, represents interspecies variability to toxicity of a toxicant. However, by not accounting for inter-test variability properly, the interpretation of the usual HC5 estimate does not align with the theoretical statistical model structure which is used. In many cases, correcting for this will not seriously alter decisions made; however, it will allow for improved quantification that can only serve to benefit risk assessment. Furthermore, it has been noted that SSDs lack ecological interpretability [25, 42, 41]; ignoring the issue of inter-test variability would only further undermine interpretability.

Although the present study focused on acute data because of their prevalence in ecotoxicity databases, chronic data are generally required by regulators for intermediate and higher tier risk assessments and for environmental standards, e.g. [12]. The NOEC, which is highly criticised by environmental statisticians

working in the field of ecotoxicology [43, 44], may be incompatible with the inter-test variability model here due to its lack of statistical robustness. A more radical approach may be to use more sophisticated models with concentration-effect data, such as that proposed by Fox [31], whereby arbitrary chronic endpoints, such as the NOEC, are replaced by modelled values. The models could, in principle, be augmented to account for some other sources of variation.

An estimate of a homogeneous inter-test standard deviation was determined to be approximately 0.47–0.51 on $\log_{10}$ μg/L concentration scale. This equates to a factor (or fold-difference) of about 3. In addition to the frequentist and Bayesian estimates being concordant, qualitative empirical analysis suggested the homogeneity assumption was reasonable.

The homogeneity assumption is the simplest model for inter-test variability. Our prior justification for starting with a parsimonious model of inter-test variability was two-fold: (1) the state of the science, namely the SSD, is itself a very simple model in reflection of the lack of available ecotoxicity data; and (2) the number of chemical-species combinations with more that one measurement was small in the database used in the present study, as shown by the grey line graph overlaid on Figure 1. More sophisticated models could be considered. For example, inter-test variability could be made to depend on taxonomic and/or chemical groups. Unfortunately, without additional data, it would not be possible to estimate all the parameters. Moreover, if certain species [groups] are typically more sensitive than others, poor fit of the SSD model may be exacerbated through the process of parameter leveraging, leading to erroneous inferences. In the interests of gaining an initial handle on the magnitude of inter-test variability and its consequences, the homogeneity model is clearly preferable as an initial step forward. Furthermore, the present study does not consider the appropriateness of the standard SSD model, a topic about which there is a a lot of on-going research [25, 40].

Exploration of the consequences of ignoring inter-test variability showed that for many chemicals, the median HC5 would be underestimated — equivalent to being over-conservative — for the majority of risk assessments. The magnitude of this difference is unlikely to be sufficient to radically reverse decisions based on the existing methods whereby inter-test variation is ignored. However, in general the importance will be proportional to the measure of the underlying true interspecies variation. EFSA [10] showed this can vary substantially between species taxonomic groups, and therefore the impact of inter-test variability may be more pronounced for some species communities.

There are two important differences between the two analyses done here. Firstly, and fundamentally, there are differences between the two models describing the SSD, namely the *status quo* model (each [aggregated] species log-toxicity value is a random observation from a normal distribution) and the *hierarchical* model (each unknown species true log-toxicity value is a random sample from a normal distribution but we can make multiple observations with error). Secondly, since the underlying models are different, matching of

prior distributions is not a trivial concept. Nonetheless, the frequentist estimator is equivalent to a Bayesian model which assumes the Jeffreys prior distribution for ψ_i ($\pi(\psi_i) \propto \psi_i^{-1}$) where the $\mu_{ij} = \bar{y}_{ij}$. That is, the inter-test variability model is discarded and the 'true' log toxicity values are replaced by the aggregated (geometric mean) toxicity values for chemical-species combination (i, j), denoted $\bar{y}_{ij}$. The more general hierarchical model based estimator assumes a uniform prior distribution ($\pi(\psi_i) \propto 1$). It can be shown that the Jeffreys prior cannot be assigned to the ψ_i parameters in the hierarchical model due to posterior probability density function being improper [37]. Consequently, posterior distributions will be different, however this will typically only be noticeable for small n_i.

It has been made clear elsewhere [41, 14] that censored data are valid for ecotoxicological risk assessment, including the quantification of inter-test variability. The standard practices of fitting SSDs and estimating HC5s (e.g. method-of-moments, look-up tables based on prior derived asymptotic theory, and graphical regression models) do not facilitate or readily include the tools necessary by risk assessors to incorporate censored data values into their analyses. Despite the existence of proposals, which are relatively difficult to implement, for augmenting the existing tools, the Bayesian counterpart is clearly preferable since it straightforwardly handles censored data whilst coherently measuring uncertainty in hierarchical models. The Bayesian paradigm also offers a rich framework to include subjective prior knowledge which will undoubtedly allow experts with specialities in specific chemical groups and species to come together to reduce uncertainty quantitatively whilst providing a transparent mechanism with which to examine expert judgements *a posteriori*. A grand model would also seek to include correlation structure for the underlying 'true' species toxicity values, such as that implemented in the US EPA Interspecies Correlation Estimation programme [18]. Such an exercise to incorporate all these features is beyond the scope of the present study which intends only to naturally extend the basic normal SSD model to include inter-test variation and to serve as a platform for risk assessors to build upon.

5 Conclusions

Based on an acute toxicity subset of quality controlled ecotoxicity database, the standard deviation of inter-test variability was quantified to be approximately 0.47–0.51 $\log_{10}$ μg/L on the log (base 10) concentration scale, equivalent to a factor (or 3-fold difference) of about 3. It is a risk management decision as to whether this constitutes a value of concern, however it is a source of uncertainty nonetheless and should be discussed in risk assessments since it will only serve to compound with other sources of uncertainty. In many assessments, accounting for inter-test variability will lead to larger (or equivalently, relatively less conservative) estimates of the HC5 — a fundamental component in risk characterisation under the REACH

guidance document — compared to those derived from current methodology.

Supporting Information

Supporting Information 1: A cleaned version of the ecotoxicity database used to quantify the inter-test variance parameter as described in De Zwart [24]. Supporting Information 2: A mathematical description of the Bayesian analysis of measurement including the R script used to perform the Bayesian calculations in this paper.

Acknowledgements

We thank the following people who provided careful reviews of the research: Stuart Marshall (Unilever), Oliver Price (Unilever), Andy Hart (The Food and Environment Research Agency), Mick Hamer (Syngenta), Mathijs Smit (Statoil ASA), Peter Chapman (Tecsolve) and Malyka Galay-Burgos (ECETOC). We also thank Unilever and Statoil ASA who funded Hickey's postdoctoral fellowship and the research conducted at Durham University, and ECETOC who helped coordinate the project. We are grateful to the two anonymous reviewers and editor for their detailed comments.

References

[1] ECHA (European Chemicals Agency). May, 2008a. Guidance on information requirements and chemical safety assessment. Part E: Risk Characterisation.

[2] EC (European Commission). 2002. Guidance document on aquatic ecotoxicology in the context of the directive 91/414/EEC. SANCO/3268/2001, 4:62.

[3] ICMM (International Council on Metals and Mining). 2007a. Metals Environmental Risk Assessment Guidance (MERAG. Fact Sheet 1: Risk Characterisation General Aspects.

[4] ICMM (International Council on Metals and Mining). 2007b. Metals Environmental Risk Assessment Guidance (MERAG. Fact Sheet 3: Effects Assessment - Data Compilation, Selection and Derivation of PNEC Values for the Risk Assessment of Different Environmental Compartments (Water, STP, Soil, Sediment).

[5] Stephan CE, Mount DI, Hansen DJ, Gentile JH, Chapman GA, Brungs WA. 1985. Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. Report No. PB 85-227049. U.S. Environmental Protection Agency, Office of Research and Development, Duluth, MN.

[6] US EPA (U.S. Environmental Protection Agency). 1998. Guidelines for ecological risk assessment; notice. *Federal Register*, 63:26846–26924.

[7] US EPA (U.S. Environmental Protection Agency). 2004. Overview of the ecological risk assessment process in the office of pesticide programs. Office of Prevention, Pesticides and Toxic Substances, Washington, DC.

[8] CCME (Canadian Council of Ministers of the Environment). 2007. A protocol for the derivation of water quality guidelines for the protection of aquatic life. Winnipeg, MB: CCME.

[9] ANZECC and ARMCANZ (Australian and New Zealand Environment and Conservation Council and Agriculture and Resource Management Council of Australia and New Zealand). 2000. Australian and New Zealand guidelines for fresh and marine water quality. National Water Quality Management Strategy Paper No. 4. ANZECC and ARMCANZ, Canberra.

[10] EFSA (European Food Safety Authority). 2005. Opinion of the scientific panel on plant health, plant protection products and their residues on a request from EFSA related to the assessment of the acute

and chronic risk to aquatic organisms with regard to the possibility of lowering the uncertainty factor if additional species were tested. *The EFSA Journal*, 301:1–45.

[11] Posthuma L, II Suter GW, Traas TP. eds. 2002. *Species sensitivity distributions in ecotoxicology.* Boca Raton: Lewis Publishers.

[12] ECHA (European Chemicals Agency) July, 2008b. Guidance for the implementation of REACH: Guidance on information requirements and chemical safety assessment. Chapter R.10: Characterisation of dose-response for environment.

[13] Aldenberg T, Jaworska JS. 2000. Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions. *Ecotoxicology and Environmental Safety*, 46:1–18.

[14] Hickey GL, Kefford BJ., Dunlop JE, Craig PS. 2008. Making species salinity sensitivity distributions reflective of naturally occurring communities: using rapid testing and Bayesian statistics. *Environmental Toxicology and Chemistry*, 27:2403–2411.

[15] Craig PS, Hickey GL, Hart A, Luttik R. 2012. Species non-exchangeability in probabilistic ecotoxicological risk assessment. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 175:243–262.

[16] Duboudin C, Ciffroy P, Magaud H. 2004. Effects of Data Manipulation and Statistical Methods On Species Sensitivity Distributions. *Environmental Toxicology and Chemistry*, 23:489–499.

[17] EFSA (European Food Safety Authority). 2006. Opinion of the scientific panel on plant health, plant protection products and their residues on a request from EFSA related to the aquatic risk assessment for cyprodinil and the use of a mesocosm study in particular. *The EFSA Journal*, 329:1–77.

[18] Raimondo S, Vivian DN, Barron MG. 2010b. Web-based interspecies correlation estimation (Web-ICE) for acute toxicity: user manual. Version 3.1. EPA/600/R-10/004. Office of Research and Development, United States Environmental Protection Agency: Gulf Breeze, FL.

[19] Baird DJ, Barber I, Bradley M, Calow P, Soares, A. 1989. The *Daphnia* bioassay: a critique. *Hydrobiologia*, 188/189:403–406.

[20] Raimondo S, Jackson CR, Barron MG 2010a. Influence of taxonomic relatedness and chemical mode of action in acute interspecies estimation models for aquatic species. *Environmental Science and Technology*, 44:7711–7716.

[21] Fairbrother A. 2008. Risk management safety factor. In: Jørgensen, SE, Faith BD. eds. *Encyclopaedia of Ecology*, Vol. 4. Elsevier Publishing: Oxford, pp 3062–3068.

[22] Raimondo S, Mineau P, Barron MG. 2007. Estimation of chemical toxicity in wildlife species using interspecies correlation models. *Environmental Science and Technology*, 41:5888–5894.

[23] ECHA (European Chemicals Agency). July, 2008c. Guidance for the implementation of REACH: Guidance on information requirements and chemical safety assessment. Chapter R.19: Uncertainty analysis.

[24] De Zwart D. 2002. Observed regularities in SSDs for aquatic species. In: Posthuma L, II Suter GW, Traas TP. eds. *Species sensitivity distributions in ecotoxicology.* Boca Raton: Lewis Publishers, pp 133–154.

[25] Newman MC, Ownby DR, Mezin LCA, Powell DC, Christensen TRL, Lerberg SB, Anderson BA. 2000. Applying species sensitivity distributions in ecological risk assessment: assumptions of distribution type and sufficient numbers of species. *Environmental Toxicology and Chemistry*, 19:508–515.

[26] Rice JA. 1994. *Mathematical statistics and data analysis.* 2nd ed. Boca Raton: Duxbury Press.

[27] Luttik R, Aldenberg T. 1997. Extrapolation factors for small samples of pesticide toxicity data: special focus on LD50 values for birds and mammals. *Environmental Toxicology and Chemistry*, 16:1785–1788.

[28] Aldenberg T, Luttik R. 2002. Extrapolation Factors for tiny toxicity data sets from species sensitivity distributions with known standard deviation. In: Posthuma L, II Suter GW, Traas TP. Eds. *Species sensitivity distributions in ecotoxicology.* Boca Raton: Lewis Publishers, 103–118.

[29] Grist EPM, O'Hagan A, Crane M, Sorokin N, Sims I, Whitehouse P. 2006. Bayesian and time-independent species sensitivity distributions for risk assessment of chemicals. *Environmental Science and Technology*, 40:395–401.

[30] Hickey GL. 2010. Ecotoxicological risk assessment: developments in PNEC estimation. Ph.D. Thesis, Durham University.

[31] Fox D. 2010. A Bayesian approach for determining the no effect concentration and hazardous concentration in ecotoxicology. *Ecotoxicology and Environmental Safety*, 73:123–131.

[32] Hayashi TI, Kashiwagi N. 2010a. A Bayesian method for deriving species-sensitivity distributions: selecting the best-fit tolerance distributions of taxonomic groups. *Human and Ecological Risk Assessment*, 16:251–263.

[33] Hayashi TI, Kashiwagi N. 2010b. A Bayesian approach to probabilistic ecological risk assessment: risk comparison of nine toxic substances in Tokyo surface waters. *Environmental Science and Pollution Research*, 16:1–11.

[34] Warren-Hicks WJ , Hart A. 2010. *Application of uncertainty analysis to ecological risks of pesticides.* Society of Environmental Toxicology and Chemistry: Brussels.

[35] Aldenberg T, Jaworska JS, Traas TP. 2002. Normal species sensitivity distributions and probabilistic ecological risk assessment. In: Posthuma L, II Suter GW, Traas, TP. eds. *Species sensitivity distributions in ecotoxicology.* Boca Raton: Lewis Publishers, pp 49–102.

[36] Hickey GL, Craig PS, Hart A. 2009. On the application of loss functions in determining assessment factors for ecological risk. *Ecotoxicology and Environmental Safety*, 72:293–300.

[37] Gelman A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:515–533.

[38] Aldenberg T, Slob W. 1993. Confidence limits for hazardous concentrations based on logistically distributed NOEC toxicity data. *Ecotoxicology and Environmental Safety*, 25:48–63.

[39] Shao Q. 2000. Estimation for hazardous concentrations based on NOEC toxicity data: an alternative approach. *Environmetrics*, 11:583–595.

[40] Van Straalen, NM 2002. Threshold models for species sensitivity distributions applied to aquatic risk assessment for zinc. *Environmental Toxicology and Pharmacology*, 11:167–172.

[41] Kefford BJ, Palmer CG, Jooste S, Warne M, Nugegoda D. 2005. 'What is it meant by 95% of species'? An argument for the Inclusion of rapid tolerance testing. *Human and Ecological Risk Assessment*, 11:1025–1046.

[42] Forbes VE Calow P. 2002. Species sensitivity distributions revisited: a critical appraisal. *Human and Ecological Risk Assessment*, 8:1625–1640.

[43] Chapman PM, Caldwell RS, Chapman PF. 1996. A warning: NOECs are inappropriate for regulatory use. *Environmental Toxicology and Chemistry*, 15:77–79.

[44] Pires AM, Branco JA, Picado A, Mendonça E. 2002. Models for the estimation of a 'no effect concentration'. *Environmetrics*, 13:15–27.

Figure captions

Figure 1: Boxplots of standard deviations [left axis] of log (base 10) acute toxicity values for chemical-species combinations (i, j) with K_{ij} records [horizontal axis]. Red horizontal dashed line is the pooled standard deviation, s_{pooled}, calculated from Eqn. 1. Translucent blue band indicates a 95% probable interval based on the assumption of normality and $\sigma = s_{\text{pooled}}$. Grey line gives the number of records (using the vertical axis on the right-hand side) in the robust subset of the database for each K_{ij} bin.

Figure 2: Posterior kernel density function for σ^2 based on 4000 samples drawn from the joint posterior distribution (Eqn. 3). See Supporting Information for technical description.

Figure 3: Plot of $\log_{10}$(HC5) estimates for 339 chemicals (i) adjusted to take into account inter-test variability [horizontal axis]; (ii) calculated using the usual methodology which does not account for inter-test variability [vertical axis]. Left panel: using all data in the robust acute effects database subset. Right panel: using only pointwise data in the robust acute effects database subset. N.B. the unadjusted estimates are based only on pointwise data. The legend provides an indication of the number of species with pointwise measurements tested for each chemical i.

Figure 1

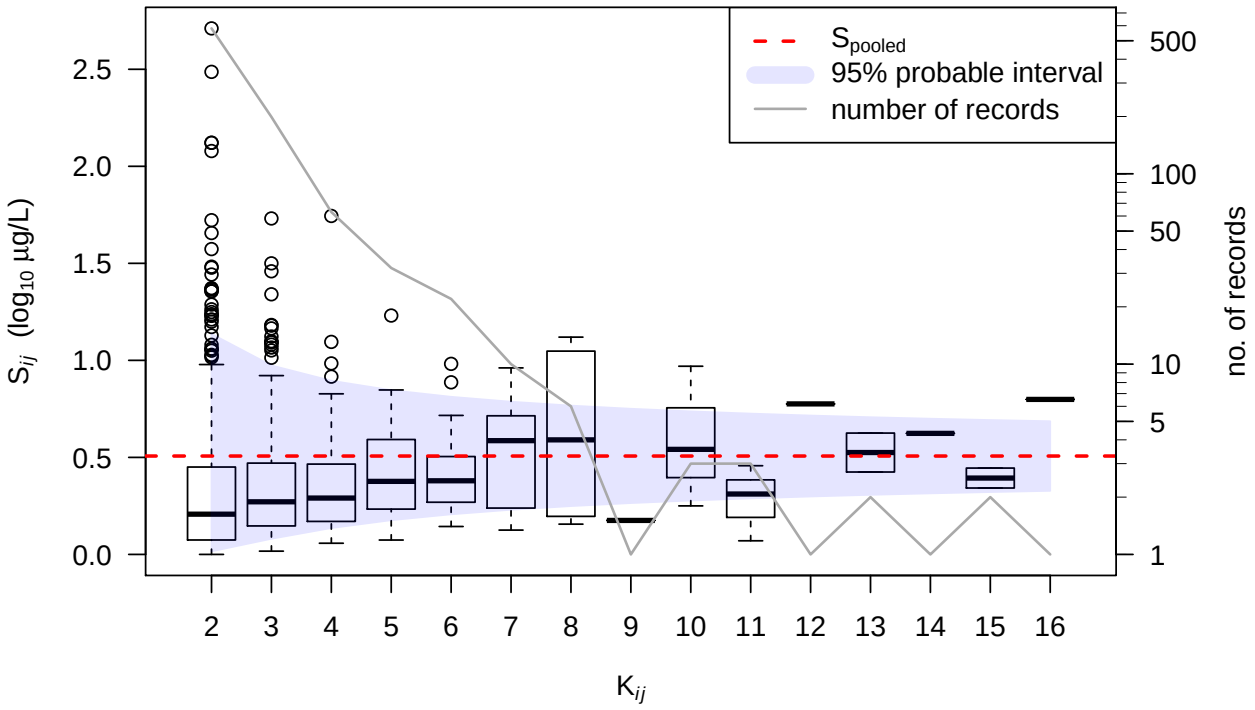

S_pooled
95% probable interval
number of records
S_ij (log10 µg/L)
no. of records
K_ij
0.0
0.5
1.0
1.5
2.0
2.5
1
5
10
50
100
500
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Figure 2

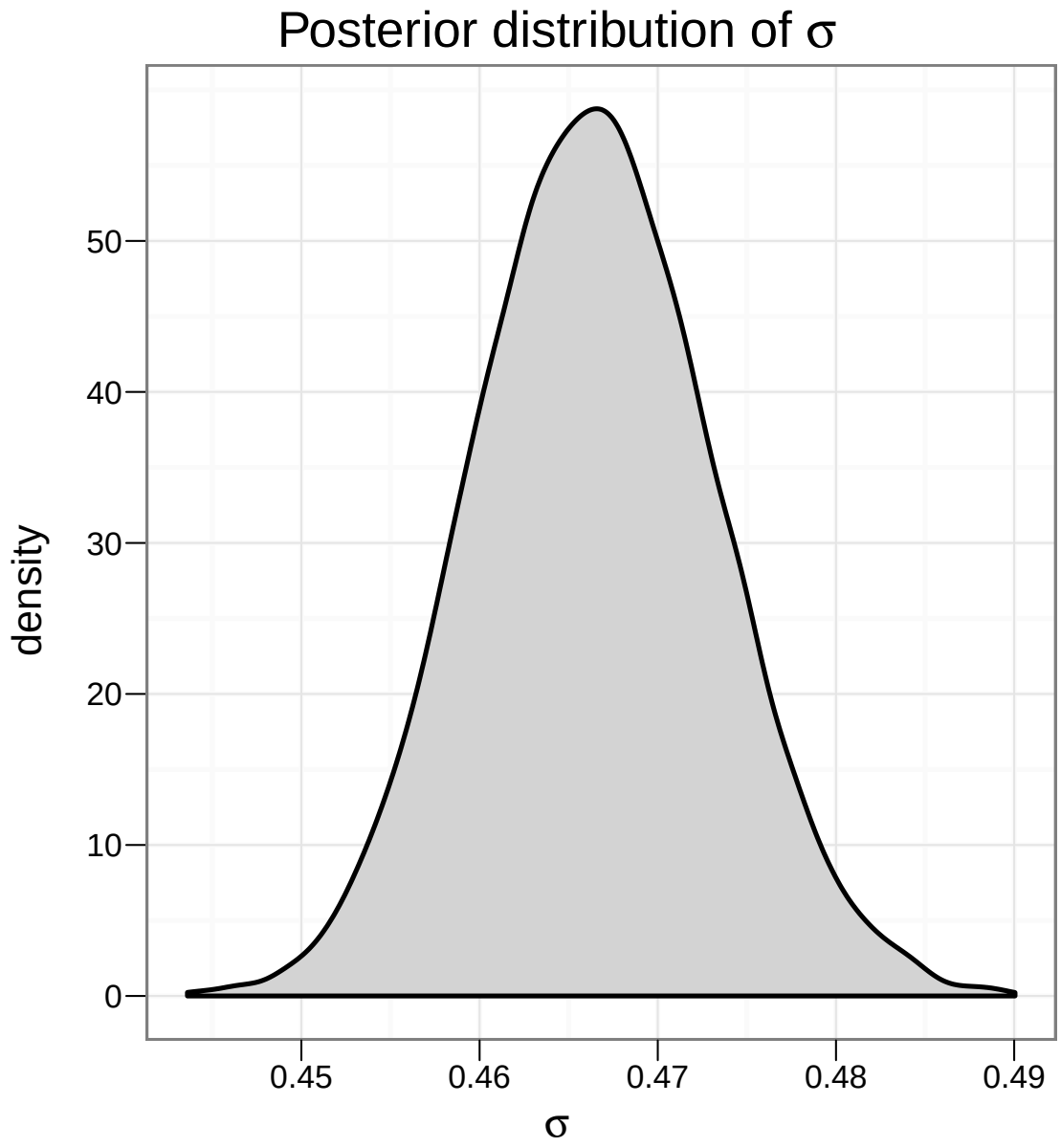
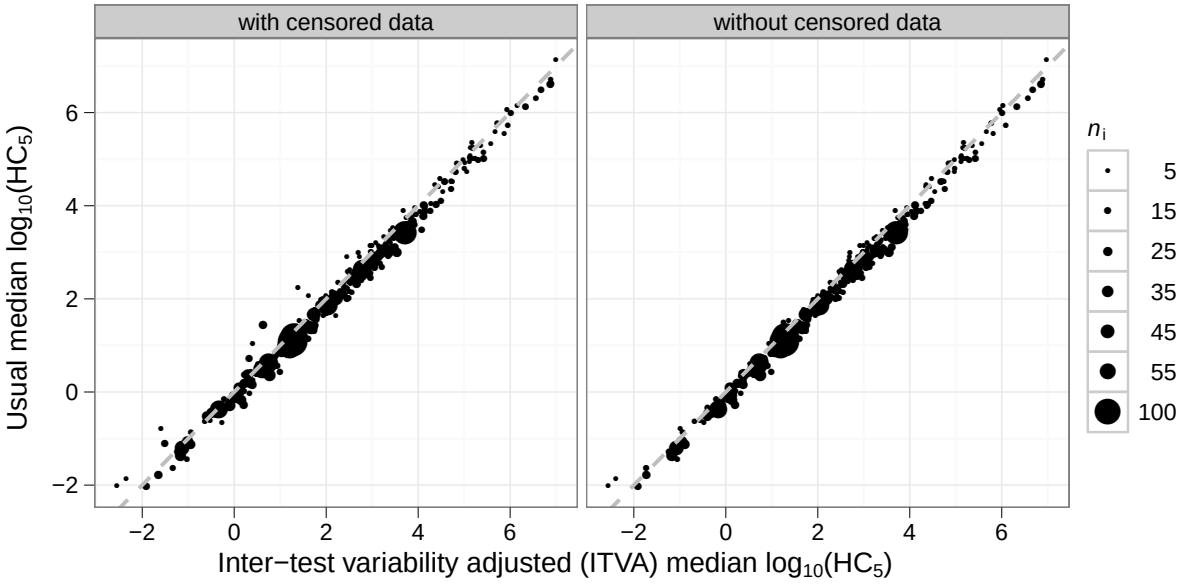

Posterior distribution of σ
density
50
40
30
20
10
0
0.45
0.46
0.47
0.48
0.49
σ

Figure 3


with censored data
without censored data
Usual median log10(HC5)
Inter−test variability adjusted (ITVA) median log10(HC5)
6
4
2
0
−2
−2
0
2
4
6
ni
5
15
25
35
45
55
100

	EC_{50}	LC_{50}	NOEC	Total
A	5	12	0	17
I	79	467	33	579
R	944	948	35	1927
P	9777	15963	1852	27592
L	37	61	156	254
Total	10842	17451	2076	30369

Table 1: Summary of the ecotoxicity database according to the endpoint and datapoint type. A = approximate (i.e. $\approx x$); I = interval censored (i.e. $> x_1$ and $< x_2$); R = right censored (i.e. $> x$); L = left censored (i.e. $< x$); P = pointwise (i.e. $= x$). The final column and row give the total number of observations over observational status and endpoint respectively. EC_{50} = median effect concentration; LC_{50} = median lethal effect concentration; NOEC = no observed effect concentration.

Supporting Information for:
On the quantification of inter-test variability in ecotoxicity data with application to species sensitivity distributions

G. L. Hickey[1], P. S. Craig[1], R. Luttik[2] & D. De Zwart[2]
[1]*Department of Mathematical Sciences, Durham University, UK*
[2]*RIVM, Bilthoven, The Netherlands*

February 23, 2012

A Technical Details of Bayesian Analysis

A.1 Derivation of the Posterior Distribution

The simplest hierarchical model for SSDs which incorporates measurement error is

$$\begin{aligned} y_{ijk} \,|\, \mu_{ij}, \sigma^2 &\sim N(\mu_{ij}, \sigma^2); \text{ and} \\ \mu_{ij} \,|\, \alpha_i, \psi_i &\sim N(\alpha_i, \psi_i^2) \end{aligned} \tag{1}$$

where y_{ijk} is the k-th $(= 1, \ldots, K_{ij})$ log (base 10) toxicity value for chemical i $(= 1, \ldots, N)$ tested on species j. For convenience, define J_i to be the set of species tested with chemical i and $\mathbf{Y}$ to be the entire database of measured log toxicity measurements. The (hyper-)parameters are assigned prior distributions as follows:

$$\begin{aligned} \pi(\sigma^2) &\sim \sigma^{-2} \text{ for } \sigma^2 > 0; \\ \pi(\alpha_i) &\sim 1 \text{ independently for each } \alpha_i \in \mathbb{R},\ i = 1, \ldots, N; \text{ and} \\ \pi(\psi_i) &\sim 1 \text{ independently for each } \psi_i > 0,\ i = 1, \ldots, N. \end{aligned}$$

Conditional on *observing* log toxicity values, the full-data likelihood function, which is the probability of observing the data but as a function of the model parameters, is given as

$$\ell(\boldsymbol{\mu}, \sigma^2) = \prod_{i=1}^{N} \prod_{j \in J_i} \prod_{k=1}^{K_{ij}} (2\pi)^{-\frac{1}{2}} \sigma^{-1} \exp\left\{ -\frac{1}{2\sigma^2} (y_{ijk} - \mu_{ij})^2 \right\},$$

where $\boldsymbol{\mu}$ is the vector of 'true' toxicity values μ_{ij} for all relevant chemical-species combinations (i, j). This can be simplified for analytical tractability as

$$\ell(\boldsymbol{\mu}, \sigma^2) \propto \prod_{i=1}^{N} \prod_{j \in J_i} \sigma^{-K_{ij}} \exp\left\{ -\frac{K_{ij}}{2\sigma^2} (\mu_{ij} - \bar{y}_{ij})^2 \right\} \exp\left\{ -\frac{(K_{ij} - 1)}{2} \frac{s_{ij}^2}{\sigma^2} \right\},$$

where $\bar{y}_{ij}$ and s^2_{ij} are the sample mean and variance of the log toxicity values for chemical-species combination (i, j).

The μ_{ij} values are 'nuisance' parameters in this analysis and would ordinarily be integrated out of the density function. However, the later complication of censored data requires us to work with the full posterior. The posterior distribution of μ, σ^2 and the hyper-parameters, $\alpha_1, \ldots, \alpha_N, \psi_1, \ldots, \psi_N$, can then be determined using Bayes' rule:

$$\pi(\boldsymbol{\mu}, \sigma^2, \alpha_1, \ldots, \alpha_N, \psi_1, \ldots, \psi_N \,|\, \mathbf{Y}) \quad \propto \quad \ell(\boldsymbol{\mu}, \sigma^2)\pi(\boldsymbol{\mu} \,|\, \alpha_1, \ldots, \alpha_N, \psi_1, \ldots, \psi_N)\pi(\sigma^2) \prod_{i=1}^{N} \pi(\alpha_i)\pi(\psi_i), \quad (2)$$

where each μ_{ij} conditional on α_i and ψ_i independently follows the distribution given by Eqn. 1.

A.2 Sampling

In order to sample from this posterior distribution, a block Gibbs Markov chain Monte Carlo (MCMC) sampler was written. The Gibbs sampler requires the posterior distribution of each parameter conditional on all the others in the model. With these distributions, starting from a "best" guess for the parameters, we cycle through them sampling each block of random variables one at a time based on the most recent version of the other parameters. All the conditional distributions described here belong to standard families (e.g. Gaussian) and therefore sampling from them is trivial once the location, scale and shape parameters are analytically determined. Further details of MCMC techniques can be found in Gelman et al. [1].

Derivation of the conditional distributions follows on straightforwardly from the decomposition in Eqn. 2; here we list them.

A.2.1 Conditional distributions: $\boldsymbol{\mu}$

$$\mu_{ij} \,|\, \alpha_i, \psi_i, \sigma^2; \mathbf{Y} \sim N\left(\left(\frac{K_{ij}\bar{y}_{ij}}{\sigma^2} + \frac{\alpha_i}{\psi_i^2}\right)\omega_{ij}, \omega_{ij}\right),$$

where,

$$\omega_{ij} = \left(\frac{K_{ij}}{\sigma^2} + \frac{1}{\psi_i^2}\right)^{-1}.$$

A.2.2 Conditional distributions: $\psi_1, \ldots, \psi_N, \alpha_1, \ldots, \alpha_N$

$$\psi_i^{-2} \,|\, \boldsymbol{\mu}, \sigma^2; \mathbf{Y} \sim \Gamma\left(\frac{n_i - 2}{2}, \sum_{j \in J_i} \frac{(\mu_{ij} - \bar{\mu}_i)^2}{2}\right),$$

$$\alpha_i \,|\, \boldsymbol{\mu}, \psi_i, \sigma^2; \mathbf{Y} \sim N\left(\bar{\mu}_i, \psi_i^{-2}/n_i\right).$$

where n_i is the number of unique species tested with chemical i, i.e. the cardinality of the set J_i, and $\bar{\mu}_i = (\mu_{i1} + \cdots + \mu_{in_i})/n_i$.

A.2.3 Conditional distribution: σ^{-2}

$$\sigma^{-2} \,|\, \boldsymbol{\mu}, \alpha_1, \ldots, \alpha_N, \psi_1, \ldots, \psi_N; \mathbf{Y} \sim \Gamma\left(\tfrac{1}{2}\sum_{i=1}^{N}\sum_{j \in J_i} K_{ij}, \tfrac{1}{2}\left(K_{ij}(\mu_{ij} - \bar{y}_{ij})^2 + (K_{ij} - 1)s^2_{ij}\right)\right).$$

A.3 Censored Data

The posterior distribution calculations above are reliant on the database of measurements, $\mathbf{Y}$, all being observed. Frequently laboratory measurements will yield censored measurements such that $y_{ijk} \in (L_{ijk}, U_{ijk})$ where L_{ijk} and U_{ijk} define the lower and upper bounds of the measurement value respectively. The type of censoring depends on the values of L_{ijk} and U_{ijk} as described in the table below.

L_{ijk}	U_{ijk}	Censoring
finite	finite	interval
finite	∞	right
$-\infty$	finite	left

Write $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{cens}})$, where $\mathbf{Y}_{\text{obs}}$ is the collection of observed measurements and $\mathbf{Y}_{\text{cens}}$ is the collection of (unknown) censored measurements with corresponding (known) intervals $(\mathbf{L}, \mathbf{U})$. Then the posterior distribution of all the unknown parameters and $\mathbf{Y}_{\text{cens}}$ conditional on $\mathbf{Y}_{\text{obs}}$, $\pi(\mathbf{Y}_{\text{cens}}, \boldsymbol{\mu}, \alpha_1, \ldots, \alpha_N, \psi_1, \ldots, \psi_N, \sigma^2 \,|\, \mathbf{Y}_{\text{obs}})$, has the same form as the right-hand side of Eqn. 2. Hence, the Gibbs sampler can be augmented with the additional conditional distributions for all data $y_{ijk} \in \mathbf{Y}_{\text{cens}}$:

$$y_{ijk} \,|\, \boldsymbol{\mu}, \sigma^2 \sim N(\mu_{ij}, \sigma^2)$$

restricted to $L_{ijk} \leq y_{ijk} \leq U_{ijk}$. It is useful to exploit the probability integral transform to generate this sample:

Step 1. Set

$$P_{L_{ijk}} = \Phi\left(\frac{L_{ijk} - \mu_{ij}}{\sigma}\right) \text{ and } P_{U_{ijk}} = \Phi\left(\frac{U_{ijk} - \mu_{ij}}{\sigma}\right),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Step 2. Randomly generate $U_{ijk} \sim U(P_{L_{ijk}}, P_{U_{ijk}})$ where $U(a, b)$ is the uniform distribution with support on $[a, b]$.

Step 3. Set $y_{ijk} = \mu_{ij} + \Phi^{-1}(U_{ijk})\sigma$

As per the model (hyper-)parameters, it is necessary to initialise the Markov chain at some possible value which satisfies the constraints of the censoring.

A.4 Implementation

A technical issue arises in the resulting posterior distribution regarding whether its normalisation constant is finite, stemming from the variance component parameters $\psi_1, \ldots, \psi_N$. However, by restricting the minimum sample size of the number of distinct species tested with each chemical to $n_i \geq 3$, the issue is resolved [2]. It is conceivable that the heavy right tail deriving from the typically small sample sizes found in ecotoxicological risk assessment will influence posterior inferences. This is of particular importance since estimates of the hazardous concentration to a fixed proportion of species (i.e. the HC_p) are a function of ψ_i. Based on heuristic suggestions in Gelman [2], we therefore restrict sample sizes to $n_i \geq 5$.

The Metropolis-within-Gibbs sampler was programmed in `R` (`http://www.r-project.org/`) [3]. The code is provided in the next section. After a burn-in period of 5,000 samples (to reach stationarity of the

chain), 10,000 samples of the random variables were generated with a thinning rate of 50 (i.e. only every 50-th sample was kept; the rest discarded) to remove the presence of serial correlation. In Fig. 1 we show the autocorrelation plot and a partial time series plot of the σ parameter sample, which are two diagnostic tools used to assess convergence properties.

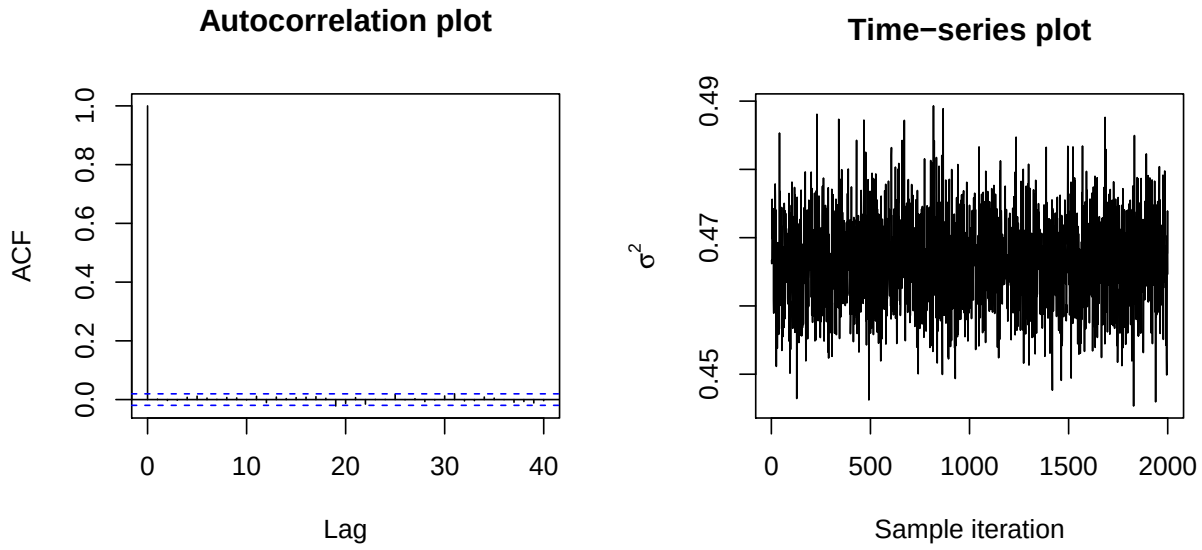


Figure 1: Diagnostic tools to assess (part of) the convergence property of MCMC sample with respect to the parameter σ. Left panel: the autocorrelation function. Right panel: time-series plot (after the burn-in period and thinning regime).

References

[1] Gelman A, Carlin JB, Stern HS, Rubin DB. 2003. *Bayesian data analysis.* 2nd ed. Chapman & Hall: Boca Raton.

[2] Gelman A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:515–533.

[3] R Development Core Team 2010. `R`: *A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. URL: `http://www.r-project.org`.

B R Code

```
##############
## PREAMBLE ##
##############

##  This code is produced by Peter Craig, Dept. of Mathematical Sciences,
##  Durham University and Graeme Hickey, NIBHI, Manchester University.
##  It is freely available for all purposes. However, the authors of the manuscript
##  assume no responsibility for any possible errors in the code.
##
##  If you have any questions, please contact: P.S.Craig@durham.ac.uk.
##
##  To being using the code, you will need to install the R statistical software
##  program and the packages listed below.

library(doBy)
library(stringr)
library(reshape)
library(ggplot2)
library(Matrix)

rivm = dget(file.choose()) # load a CSV version of the database from files
rivm = rivm[rivm$endpoint %in% c("EC50", "LC50", "NOEC"), ]
rivm = droplevels(rivm)

#-----------------------------------------------------------------

######################
## DATA MANANGEMENT ##
######################

# full robust acute (EC50/LC50)
inds1 = with(rivm,
    (endpoint == "LC50" | endpoint == "EC50") &
    (effect == "MOR" | effect == "IMM"))
inds2 = with(rivm,
    (dur.low == 2) &
    (major %in% c("CR", "IN")))
inds3 = with(rivm,
    (dur.low == 4) &
    !(major %in% c("CR", "IN")))
inds4 = !(grepl(" sp$", rivm$species) | grepl(" sp.", rivm$species))
```

```
inds5 = grepl(" ", rivm$species)
inds6 = !(rivm$major == "MI")
inds7 = !(rivm$conc.ind == "A")

acute.r = rivm[which(inds1 & (inds2 | inds3) & inds4 & inds5 & inds6 & inds7), ]
acute.r = drop.levels(acute.r)

# Robust with n >= 5 distinct species *pointwise* measurements
# (incl. censored data)
n = by(
    acute.r,
    factor(acute.r$CAS),
    function(d) length(unique(d[d$conc.ind == "P", ]$species))
)
status = (n >= 5)
acute.r2 = acute.r[acute.r$CAS %in% names(n)[status], ]
acute.r2 = drop.levels(acute.r2)

# Robust with n >= 5 distinct species *pointwise* measurements
# (not incl. censored data)
acute.r3 = acute.r2[acute.r2$conc.ind == "P", ]
acute.r3 = drop.levels(acute.r3)

##################
## MCMC SAMPLER ##
##################

## Options:
## -- data (data.frame): appropriately labelled as per acute.r2.
## -- N (integer > 1): number of MCMC samples to return.
## -- thin (integer >= 1): thinning rate.
## -- detailed (logical): return nusiance variables? Default = FALSE.

gibbs.fast = function(data, N, thin = 10, detailed = FALSE) {

  ## Add CAS:Species interaction term to each record
  data = cbind(
    ij = interaction(data$CAS, data$species, drop = TRUE),
    data[c("CAS", "species", "conc.ind", "lconc.low", "lconc.upp")]
    )

  ## Calculate 'approximate' values for y.ijk (i.e. for each measurement)
  data = transform(data,
```

```
  yapprox = ifelse(
    conc.ind %in% c("P","L"),
    lconc.low,
    ifelse(conc.ind=="U",
           lconc.upp,
           (lconc.low + lconc.upp) / 2
           )
    )
  )

## Calculate 'approximate' values for y.ij (i.e. for each CAS:Species)
data.ij = summaryBy(yapprox ~ ij, id = ~ CAS, data = data, FUN = c(length, mean))
data.ij = transform(
  data.ij,
  k = yapprox.length,
  yapprox.length = NULL
  )
k.total = sum(data.ij$k)

## Calculate 'approximate' values y.i (i.e. for each chemical SSD)
data.i = summaryBy(yapprox.mean ~ CAS, data = data.ij, FUN = c(length, mean, sd))
data.i = transform(
  data.i,
  n = yapprox.mean.length, # number of distinct species per chemical
  yapprox.mean.length = NULL
  )
data.ij$data.i.index = match(data.ij$CAS, data.i$CAS)
M2.tapply = 1.0 * outer(data.i$CAS, data.ij$CAS, "==") # no. chems x no. ij-pairs
M2.tapply = Matrix(M2.tapply, sparse = TRUE)

is.pw = data$conc.ind == "P" # Is pointwise?
interval = data[!is.pw, ] # Interval censored data
do.interval = nrow(interval) > 0 # Is there any censored data
if (do.interval)
  interval$data.index = seq(nrow(data))[!is.pw] # index of record in data
pw = data[is.pw, ] # subset of pointwise data

## Statistics for pointwise data only
pw.ij = summaryBy(
  lconc.low ~ ij,
  id = ~ CAS + species,
  data = pw,
  FUN = c(mean, sd, length)
```

```
  )
pw.ij = transform(
  pw.ij,
  ybar = lconc.low.mean, s = lconc.low.sd, k = lconc.low.length,
  lconc.low.mean = NULL, lconc.low.sd = NULL, lconc.low.length = NULL
  )
pw.ij = transform(pw.ij, ysum = k*ybar, sse = ifelse(k>1, (k-1)*s^2, 0))
pw.ij$data.ij.index = match(pw.ij$ij, data.ij$ij)

## Modify representation of interval data + add in index for unique ij-table
if(do.interval) {
  interval$data.ij.index = match(interval$ij, data.ij$ij)
  unique.interval.ij.indices = unique(interval$data.ij.index)
  M.tapply = 1.0 * outer(unique.interval.ij.indices, interval$data.ij.index
    , "==") # no. unique censored ij-pairs x no. ij-pairs
  M.tapply = Matrix(M.tapply, sparse = TRUE)
  interval[interval$conc.ind == "L", "lconc.upp"] = Inf
  interval[interval$conc.ind == "U", "lconc.low"] = -Inf
}

## Allocate memory for posterior samples
n.ij = nrow(data.ij) # no. of unique ij-pairs
n.i = nrow(data.i) # no. of chemicals
if(detailed) mu.ij.mcmc = matrix(NA, n.ij, N)
sigma.mcmc = numeric(N)
alpha.mcmc = matrix(NA, n.i, N)
psi.mcmc = matrix(NA, n.i, N)
if(do.interval) {
  n.interval = nrow(interval)
  if(detailed) y.interval.mcmc = matrix(NA, n.interval, N)
}
ysum.ij.pw = rep(0, n.ij)
ysum.ij.pw[match(pw.ij$ij, data.ij$ij)] = pw.ij$ysum

## Initial values
alpha.i = data.i$yapprox.mean.mean
psi.i = pmax(0.5, data.i$yapprox.mean.sd)
mu.ij = data.ij$yapprox.mean
sigma = with(
  subset(pw.ij, k>1),
  sqrt(sum((k-1)*s^2) / sum(k-1))
  )
```

```
## Function to sample from normal distribution with (interval) censoring
rcensnorm = function(n, low, upp, mu, sigma) {
  tophalf = low > mu
  tmp = low[tophalf]
  low[tophalf] = -upp[tophalf]
  upp[tophalf] = -tmp
  mu[tophalf] = -mu[tophalf]
  plow = pnorm(low, mu, sigma)
  pupp = pnorm(upp, mu, sigma)
  p = plow + runif(n)*(pupp-plow)
  x = qnorm(p, mu, sigma)
  x = pmin(pmax(x, low), upp)
  x[tophalf] = -x[tophalf]
  x
}

for(t in 1:(N*thin)) {

  ## Sample y.interval
  if(do.interval)
    y.interval = rcensnorm(
      n.interval,
      interval$lconc.low, interval$lconc.upp,
      mu.ij[interval$data.ij.index],
      sigma
      )

  ysum.ij = ysum.ij.pw

  if(do.interval) {
    ysum.ij.interval = as.vector(M.tapply %*% y.interval)
    ysum.ij[unique.interval.ij.indices]  =
      ysum.ij[unique.interval.ij.indices] + ysum.ij.interval
  }

  ybar.ij = ysum.ij/data.ij$k

  ## Sample sigma
  sse.pw = sum(pw.ij$sse + pw.ij$k*(pw.ij$ybar-ybar.ij[pw.ij$data.ij.index])^2)
  if(do.interval) {
    sse.interval = sum((y.interval-ybar.ij[interval$data.ij.index])^2)
    sse.total = sse.pw + sse.interval
  } else sse.total = sse.pw
```

```
  total.variation = sse.total + sum(data.ij$k * (ybar.ij-mu.ij)^2)
  sigma = 1 / sqrt(rgamma(1, k.total/2, total.variation/2))

  ## Sample mu.ij
  prec.alpha.ij = 1/psi.i[data.ij$data.i.index]^2
  alpha.ij = alpha.i[data.ij$data.i.index]
  prec.ybar.ij = data.ij$k / sigma^2
  prec.mu.ij = prec.alpha.ij + prec.ybar.ij
  E.mu.ij = (prec.alpha.ij * alpha.ij + prec.ybar.ij * ybar.ij) / prec.mu.ij
  mu.ij = rnorm(length(ybar.ij), E.mu.ij, 1/sqrt(prec.mu.ij))

  ## Sample psi.i
  mubar.i = as.vector((M2.tapply %*% mu.ij)) / data.i$n
  sse.mubar.i = as.vector(M2.tapply %*% ((mu.ij - mubar.i[data.ij$data.i.index])^2))
  df.i = data.i$n-2
  psi.i = 1 / sqrt(rgamma(length(mubar.i), df.i/2, sse.mubar.i/2))

  ## Sample alpha.i
  alpha.i = rnorm(length(mubar.i), mubar.i, psi.i / sqrt(data.i$n))

  ## Save the results
  if (t %% thin == 0) {
    if(detailed) mu.ij.mcmc[ , t/thin] = mu.ij
    sigma.mcmc[t/thin] = sigma
    if(do.interval && detailed) y.interval.mcmc[ , t/thin] = y.interval
    alpha.mcmc[ , t/thin] = alpha.i
    psi.mcmc[ , t/thin] = psi.i
  }

}

## Output
res = list(
  data = data,
  data.ij = data.ij,
  pw.ij = pw.ij,
  N = N,
  alpha = t(alpha.mcmc),
  psi = t(psi.mcmc),
  sigma = sigma.mcmc
  )
if(detailed) res$mu.ij = t(mu.ij.mcmc)
```

```
  if(do.interval) {
    res$interval = interval
    if(detailed) res$y.interval = t(y.interval.mcmc)
  }

  res

}

#-------------------------------------------------------------------------

########################
## POSTERIOR ANALYSIS ##
########################

## Extract required MCMC chains

out.c = gibbs.fast(acute.r2, N = 5000, thin = 100)
alpha.c = out.c$alpha
psi.c = out.c$psi
sigma.c = out.c$sigma

out.p = gibbs.fast(acute.r3, N = 5000, thin = 100)
alpha.p = out.p$alpha
psi.p = out.p$psi
sigma.p = out.p$sigma

## Calculate posterior median log(HC5)s

delta.c = alpha.c - qnorm(0.95)*psi.c
hc5.bayes.c = data.frame(
    CAS = levels(acute.r2$CAS),
    delta.tilde = apply(delta.c, 2, median),
    alpha.mean = apply(alpha.c, 2, mean),
    psi.mean = apply(psi.c, 2, mean))

delta.p = alpha.p - qnorm(0.95)*psi.p
hc5.bayes.p = data.frame(
    CAS = levels(acute.r3$CAS),
    delta.tilde = apply(delta.p, 2, median),
    alpha.mean = apply(alpha.p, 2, mean),
    psi.mean = apply(psi.p, 2, mean))
```

```
hc5.bayes = merge(hc5.bayes.c, hc5.bayes.p,
    by = "CAS", suffixes = c(".c", ".p"))

## Calculate frequentist median log(HC5)s

# Aggreagtion of acute.r3 over chemicals
acute.r3.agg = summaryBy(
    lconc.low ~ CAS + species,
    data = acute.r3,
    FUN = mean,
    keep.names = TRUE)

hc5.freq = do.call("rbind", by(
    acute.r3.agg,
    acute.r3.agg$CAS,
    FUN = function(d) {
        y = d$lconc.low;
        data.frame(
            CAS = d$CAS[1],
            n = length(y),
            ybar = mean(y),
            s = sd(y)
        )
    }
))
hc5.freq$delta.hat = with(
    hc5.freq,
    ybar - qt(0.5, n-1, qnorm(0.95)*sqrt(n))*s/sqrt(n))

## Combine into an overall summary dataframe

hc5 = merge(hc5.bayes, hc5.freq, by = "CAS")
d = melt(hc5,
    id.vars = c("CAS", "n", "delta.hat"),
    measure.vars = c("delta.tilde.c", "delta.tilde.p"))
levels(d$variable) = c("with censored data", "without censored data")

## Generate plot of log(HC5)s: Bayes-corrected vs. freq.

p = ggplot(d, aes(x = value, y = delta.hat))
p = p + geom_point(aes(size = n)) +
    xlab(expression(paste("Measurement error adjusted median log"[10], "(HC"[5], ")"))) +
    ylab(expression(paste("Usual median log"[10], "(HC"[5], ")")))
```

```
p = p + facet_grid(.~variable)
p = p + scale_size(expression(italic(n)[i]), breaks = c(seq(5, 55, 10), 100))
p = p + geom_abline(intercept = 0, slope = 1, colour = "grey", line = "dashed", size = 0.8)
p + theme_bw()

## Generate plot of posterior density of sigma

df = data.frame("sigma" = sigma.c)
q = ggplot(aes(x = sigma), data = df)
q = q + geom_density(fill = "lightgrey") + xlab(expression(sigma))
q + theme_bw(base_size = 9) +
    opts(title = expression(paste("Posterior distribution of ", sigma)))

## Display diagnostics of sigma

par(mfrow = c(1, 2))
acf(sigma.c, main = "Autocorrelation plot")
plot(ts(sigma.c), main = "Time-series plot", xlab = "Sample iteration", ylab = expression(sigma))
```